

Intégration de la dimension étymologique dans le modèle OntoLex-Lemon : le cas particulier de la lexicographie historique romane

1. Introduction

Derrière l'intitulé un peu technique de cet article se cache une interrogation centrale pour la linguistique historique romane : comment, dans le passage au numérique, peut-on donner accès aux nombreuses références et mises à jour que se font mutuellement nos ressources lexicographiques ? Cette interrogation concerne particulièrement les ouvrages de synthèse que sont par exemple le FEW, le LEI ou le DÉRom, qui rassemblent une vaste documentation décrite ailleurs et qui sont eux-mêmes continuellement cités et corrigés au sein de la littérature scientifique. Dans la version imprimée de nos ressources, ces citations et corrections sont signalées par des références aux articles concernés et par des remarques expliquant les corrections à effectuer. L'*Atlas linguistique de la Wallonie* (ALW), par exemple, corrige le FEW lorsqu'il attribue un étymon au lexème *zamté* "essaimer" (ALW 17, 99), que l'on trouve notamment dans le village belge de Francorchamps (désigné par l'abréviation « Ve37 »), alors que le FEW avait regroupé ce verbe et d'autres similaires dans les matériaux d'origine inconnue. L'ALW signale explicitement que ces formes, mentionnées dans FEW 21, 342ab s.v. *courir*, sont à déplacer dans l'article du FEW qui regroupe la descendance du lexème latin *examen*, à savoir FEW 3, 258a (voir figure 1).

G. 1. 'fait (es)saimer':⁶ f. *samé* S 1 (proprement 'essaimer'), 19, '32 l -*æ* Mo 17 | f. *šómè* D 7 (proprement 'essaimer') || 'fait (es)saimer'. *f. **zam'ter**, **cham'ter**, **chim'ter**, **chom'ter**...⁷ f. *zamté* Ve 37, 39, 41, 42; My 3; Ma 3, 9, 20; B 9, 12 | f. *zamtey* W 63 | f. *samtè* D 46 | f. *zamtî* B 7 ('fait sauver') | f. *šamté* S 19, 36, 37; Ch 26, 28; Th '2; Ni 19, 20 (ou ~ *èvōy*), 28, 112; Na '20, 23; Ph 45; H 38, 46; [...]

⁷ Tous ces verbes, regroupés FEW 21, 342ab *courir*, sont à déplacer FEW 3, 258a EXAMEN; comp. liég. **samer* 'essaimer' et pic. **déchamer* 'déguerpier'. – V. aussi DEPR.-NOP.

Figure 1 – ALW 17, 99 : étymologie du verbe *zamté*

Dans une version numérique de ces deux ressources, il est évidemment souhaitable que cette correction au FEW soit pleinement accessible, non seulement aux lecteurs de l'ALW, mais aussi – et surtout – aux lecteurs du FEW. Dans l'ALW, la référence au FEW peut être représentée par une URL envoyant le lecteur sur la page WEB du FEW en ligne (page contenant soit le début de l'article EXAMEN, soit le paragraphe exact concerné par l'insertion de ces nouveaux lexèmes). On pourrait également imaginer que la page WEB de l'article EXAMEN du FEW comporte un lien vers l'ALW, signalant cette correction, afin que le lecteur du FEW en soit averti. La mise en place d'un lien bidirectionnel entre les deux ressources

peut être automatisée (cf. Renders / Dethier / Baiwir 2015, qui propose une solution pratique permettant en outre de conserver les structures informatiques propres à chaque ressource).

Il est toutefois possible d'aller plus loin. Au lieu de créer un lien entre deux ressources (ALW et FEW), on pourrait créer un lien entre deux lexèmes, indépendamment du lieu où ils sont répertoriés et décrits. Il s'agit alors de créer une relation nouvelle entre le lexème wallon *zamté* et le lexème latin *examen*, en signalant que le second est l'étymon du premier. Précisons que ces deux lexèmes restent identifiables par leur appartenance respectivement à l'ALW et au FEW ; ils reçoivent simplement chacun une propriété de plus, spécifiant leur lien étymologique.

Cette façon de faire présente un intérêt théorique évident : en mettant à jour les données elles-mêmes, et non les ressources, elle redescend l'information au niveau précis où elle a un sens. Ce faisant, elle apporte une solution au problème de la mise à jour de nos ouvrages. Il n'est en effet plus nécessaire de se demander comment insérer les lexèmes de l'ALW au sein de l'article *EXAMEN* du FEW, au risque de déstructurer ce dernier ou de rendre sa lecture plus difficile. La mise à jour du FEW s'effectue en dehors du FEW, à un niveau indépendant de ce dernier, par la mise en place d'une relation nouvelle entre une donnée du FEW d'une part et une donnée de l'ALW d'autre part. Le fait que la mise en relation de ces deux données soit explicitée sémantiquement (l'un est l'étymon de l'autre) est la condition qui permet cette mise à jour.

C'est exactement à cela que sert le Linked Data et c'est ce que voudrait expliquer cet article. Après un bref rappel des principes qui guident le WEB sémantique (2), nous décrivons le modèle qui est actuellement proposé comme standard pour l'intégration de la lexicographie numérique au sein du Linked Data (3) et, ensuite, la façon dont nous pouvons l'exploiter pour mettre en réseau et mettre à jour les ressources de référence en linguistique historique romane (4).

2. WEB sémantique et Linked Data

Le WEB sémantique a pour objectif de décrire les ressources du WEB et leurs relations. Ces descriptions se formalisent très simplement par ce qu'on appelle un triplet RDF (pour *Resource Description Framework*), présentant deux objets reliés par une propriété. L'ensemble de ces objets et propriétés forme un graphe, au sein duquel tout objet est susceptible de servir de point d'entrée. Une des différences entre le WEB et le WEB sémantique réside dans le fait que ces objets ne sont pas nécessairement des pages WEB, mais les données elles-mêmes. Chacun des objets et propriétés est identifié de façon unique par une adresse URI (*Unique Resource Identifier*), selon les principes du Linked Data définis par Tim Berners-Lee (voir Tittel / Chiarcos 2018 pour un résumé de ces principes).

Dans notre exemple, la donnée de l'ALW *zamté* pourrait être l'objet de plusieurs descriptions RDF, explicitant notamment sa nature (« *zamté* est une forme »), son appartenance à une ressource (« *zamté* appartient à l'ALW ») ou encore sa localisation (« *zamté* est localisé à Francorchamps »). Ces descriptions, qui mettent chacune en relation

deux objets (« *zamté* », « forme », « ALW », « Francorchamps ») à l'aide de propriétés (« est un », « appartient à », « est localisé à »), constituent autant de métadonnées permettant à des programmes informatiques d'interpréter les ressources du WEB.

Ces descriptions nécessitent un vocabulaire commun, pour désigner à la fois les types (ou *classes*¹) d'objets (le type « localisation », le type « forme ») et les propriétés (« est localisé à »). Ce vocabulaire ainsi que les données décrites au moyen de ce dernier constituent ensemble une *ontologie*. Il existe de nombreuses ontologies disponibles pour toute ressource du WEB (*dbpedia* par exemple). Chaque ressource numérique qui intègre le WEB sémantique peut dès lors utiliser des ontologies existantes et, également, créer sa propre ontologie. L'utilisation de vocabulaires partagés assure l'interopérabilité des ressources et leur intégration dans le vaste réseau du WEB sémantique.

3. *OntoLex-Lemon et ses extensions*

De nombreux vocabulaires ont été développés en fonction des divers types de ressources à décrire. Les ressources lexicales disposent depuis peu d'un modèle appelé OntoLex-Lemon, initialement créé (sous l'intitulé Lemon) pour faciliter le rendu des ontologies en langues naturelles en les complétant avec des informations linguistiques (<<https://lemon-model.net>>, McCrae *et al.* 2008). Depuis 2012, OntoLex s'impose comme le modèle standard pour l'intégration dans le WEB sémantique des ressources lexicographiques.

OntoLex part du principe, nécessairement artificiel dans une structure de données en graphe, que le point d'entrée dans le réseau sémantique est l'entrée d'un article de dictionnaire. Ce point d'entrée, appelé *LexicalEntry*, peut être relié à la ressource lexicographique qui le contient (appelée *Lexicon*) ainsi qu'aux divers types d'information susceptibles d'apparaître dans un article de dictionnaire : catégorie grammaticale, formes graphiques et phonétiques, sens, etc. Il est également relié (directement ou indirectement) à un objet de type *Concept*, peu pertinent pour la modélisation d'une ressource lexicographique, mais essentiel pour l'intégration de cette ressource au sein du WEB sémantique. L'importance d'une relation entre un *Concept* et un *LexicalEntry* s'explique aussi par l'objectif initial d'OntoLex, à savoir apporter des informations linguistiques aux concepts provenant des ontologies du WEB sémantique. Le schéma ci-dessous (figure 2) représente le noyau du modèle OntoLex.

Ce noyau a été enrichi de modules complémentaires, permettant de représenter plus finement divers niveaux de description linguistique (syntaxe, sémantique, traduction, etc., cf. <<https://lemon-model.net>>, McCrae *et al.* 2008). L'ensemble permet la modélisation RDF de la plupart des ressources lexicographiques des différentes langues. Il faut y ajouter des extensions actuellement en cours d'élaboration, parmi lesquelles des propositions visant à permettre une description fine des relations qui s'établissent entre deux lexèmes sur l'axe diachronique. Après quelques tentatives visant à représenter les informations étymologiques

¹ Par souci de rendre le propos le moins technique possible, nous désignons parfois par *type d'objet* ce qu'il est convenu d'appeler *classe d'objet* dans une modélisation RDF.

présentes dans certains dictionnaires (cf. Declerck / Wandl-Vogt / Mörth 2015), un ensemble plus complet de classes et de propriétés a été élaboré par Fahad Khan, dans le but d'intégrer au sein du Linked Data la lexicographie historique et étymologique. Ces classes et propriétés permettent de décrire l'évolution diachronique d'une unité lexicale, d'une forme ou d'un sens, en prenant en compte les différents niveaux de changement linguistique (phonologique, sémantique, etc.) et l'absence de précision temporelle ou sémantique qui accompagnent souvent la description de l'étymologie-histoire d'un lexème (Khan 2018). Les propositions prévoient en outre la représentation de diverses hypothèses étymologiques ainsi que des références à la littérature secondaire, en distinguant attestation et citation (Khan / Boschetti 2018).

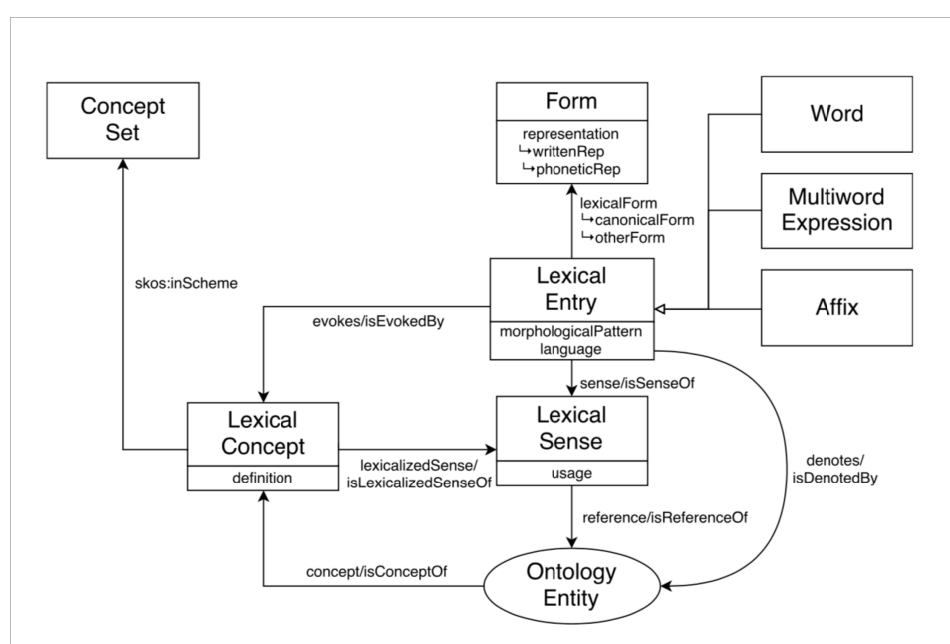


Figure 2 – Ontolex Core Model

L'extension proposée par Fahad Khan, appelée *LemonEty*, introduit dans le modèle OntoLex quelques classes d'objets et propriétés essentielles pour la représentation d'informations étymologiques. En voici les principales.

- La classe *Etymology* représente tout discours étymologique. Elle peut être reliée à une entrée de dictionnaire (*LexicalEntry*), mais aussi à une forme ou à un sens.
- Les classes *Etymon* et *Cognate* désignent des unités lexicales qui, au sein de la ressource modélisée, jouent le rôle d'étymon ou de cognat. Ces etymons et cognats étant des lexèmes eux-mêmes, susceptibles de description linguistique au même titre que les entrées du dictionnaire, ils appartiennent à la classe des objets *LexicalEntry* et peuvent recevoir les mêmes propriétés.
- La classe *EtyLink* permet la description d'une relation étymologique entre deux objets, relation qui peut être spécifiée (descendance héréditaire ou emprunt, par exemple).

- Enfin, l’extension définit les propriétés *etySource* et *etyTarget*, qui permettent de représenter respectivement la source et la cible d’une relation étymologique.

Ces ajouts au modèle OntoLex permettent d’envisager la modélisation RDF des ressources de la linguistique historique romane et la mise en réseau de ces ressources via les références d’ordre étymologique.

4. Mise en réseau des ressources de la linguistique romane

Au sein de la linguistique historique romane, le FEW et l’ALW sont à notre connaissance les premières ressources lexicographiques à avoir fait l’objet de propositions de modélisation RDF (cf. Mazziotta 2011). Le DEAF est, quant à lui, le premier dictionnaire à avoir été formalisé selon le standard OntoLex-Lemon. Nous renvoyons le lecteur à Tittel / Chiarcos (2018) pour une description détaillée de ce travail, qui a pris en compte les divers types d’information présents dans un article du DEAF (en laissant provisoirement de côté les informations de type étymologique, les propositions de Fahad Khan étant toujours en cours d’élaboration).

Une formalisation RDF ne doit pas nécessairement prendre en compte l’ensemble des informations présentes dans une ressource : seules les informations que l’on désire mettre en réseau au sein du WEB sémantique doivent être disponibles. Dès lors, une question préalable à toute tentative de formalisation RDF reste de savoir quels sont les objectifs de l’entreprise. Dans cet article, nous nous penchons sur la question de la modélisation RDF des références étymologiques et sur l’intérêt de cette modélisation pour la mise à jour d’un ouvrage tel que le FEW.

Les propositions de Khan (2018) ont été conçues pour la modélisation RDF de dictionnaires ‘classiques’, dans lesquels le discours étymologique constitue une partie structurelle de l’article. Des dictionnaires comme le TLF, le DMF ou le DEAF appartiennent à ce type de ressources. Les notices « Étymologie et histoire » du TLF peuvent par exemple être modélisées sans trop de problèmes au moyen des diverses classes et propriétés définies par Khan, de façon à identifier les étymons et cognats, les relations étymologiques et les lexèmes concernés par ces relations. L’application du modèle OntoLex dans le FEW ou le DÉRom pose en revanche quelques défis, dus à la structure particulière de ces ouvrages. L’étymon y est en effet un critère de classement, et non un type d’information parmi d’autres : ce sont des ouvrages à étymologie intégrante (et non à étymologie intégrée, cf. Buchi / Renders 2013, 660), qui rassemblent sous un étymon vedette les unités lexicales de plusieurs langues et parlers. L’étymon-entrée n’est donc pas dans ces ressources un *LexicalEntry* au sens que lui donne OntoLex. Dans ces ouvrages, il est nécessaire de considérer comme un *LexicalEntry* tout lexème appartenant aux langues qui sont l’objet de la description lexicographique, même si ce lexème ne se présente pas à proprement parler comme une entrée de dictionnaire. Ce changement de point de vue correspond à la vision du FEW comme un thesaurus (Renders 2015, 75-77). Dès lors, dans l’article EXAMEN du FEW, chacun des lexèmes rassemblés dans la documentation (par exemple le lexème fr. *essaimer*) constituera pour OntoLex un *LexicalEntry* à part entière ; l’étymon qui forme le titre de

l'article sera quant à lui représenté comme un *Etymon* et relié à ces lexèmes au moyen des propriétés de la classe *EtyLink*. À condition d'opérer ce changement de point de vue, le modèle OntoLex peut être utilisé pour représenter le lien qui unit d'une part l'étymon-entrée et d'autre part l'ensemble des lexèmes rassemblés sous cet étymon. La même remarque vaut pour le rassemblement de lexèmes sous un concept-titre, tel qu'on le trouve dans les matériaux d'origine inconnue du FEW ou encore dans les notices de l'ALW : dans l'article de l'ALW que nous avons pris pour exemple, chaque unité lexicale (comme *zamté*) constituera un *LexicalEntry* et sera relié par une propriété conceptuelle à la notion qui constitue le titre de la notice.

Une fois opéré ce changement de point de vue, il est possible de représenter à l'aide du modèle OntoLex (amélioré par les propositions de Khan 2018 et Khan / Boschetti 2018) l'ensemble des informations qui, dans ces ouvrages, sont directement associées à une unité lexicale. On accède ainsi à l'infrastructure du FEW (Büchi 1996, 5-6), tout en conservant la relation étymologique qui unit les unités lexicales à leur étymon, que ce dernier soit l'étymon-vedette de l'article ou un étymon caché. Les lexèmes de l'ALW seront de la même manière reliés à leur étymon lorsqu'il est mentionné.

Il est à noter que jusqu'ici, nous avons laissé de côté les informations qui concernent la microstructure, à savoir la façon dont le rédacteur classe les lexèmes au sein d'un article selon divers critères de regroupement (cf. Büchi 1996, 75sq.). Cette particularité structurelle, partagée par des ouvrages comme le FEW, le LEI ou le DÉRom (ainsi que l'ALW, cf. Boutier 2008), correspond à ce que nous avons appelé la dimension monographique de ces ouvrages, qui s'attachent dans leurs articles à décrire l'histoire et l'évolution d'une famille lexicale (Renders 2015, 75). Dans cette vision monographique, l'ensemble d'un article du FEW serait à placer dans la classe *Etymology*, ce qui constitue un sérieux problème de modélisation. Suivant le même principe de redescende des informations qui permet de créer un lien entre les lexèmes et l'étymon vedette d'un article du FEW, il est possible d'associer à chaque *LexicalEntry* une série d'informations concernant sa place dans la microstructure de l'article où il est décrit. Les propositions de modélisation RDF du FEW et de l'ALW (cf. Mazziotta 2011) définissent dans ce but des classes qui représentent les diverses possibilités de groupement (sémantique, étymologique, phonétique, etc.) ainsi que des propriétés permettant la hiérarchisation de ces groupements. À la condition d'actualiser ces propositions en les intégrant à celles de Khan (2018), on peut envisager de modéliser au moyen d'OntoLex la totalité des informations présentes dans un ouvrage comme le FEW, le DÉRom ou l'ALW.

Reprenons notre exemple introductif et imaginons que les articles concernés de l'ALW et du FEW ont tous deux fait l'objet d'une formalisation RDF selon les principes définis ci-dessus. Comment s'opère, au sein du WEB sémantique, la mise en réseau de nos deux ressources ? La référence que fait l'ALW au FEW nécessite de créer plusieurs relations nouvelles entre le lexème *zamté* de l'ALW et les données du FEW. La plus importante consiste à relier *zamté* à *examen* via une relation *EtyLink* (voir figure 3).

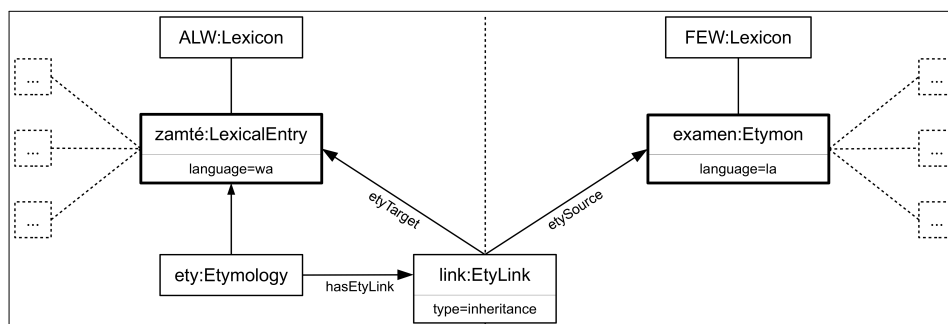


Figure 3 – Mise en réseau ALW-FEW

Cette relation suffit pour que le lexème wallon *zamté* soit relié indirectement à toutes les informations RDF associées au lexème latin *examen* et soit associé, par la même occasion, à l'ensemble des lexèmes qui ont *examen* pour étymon. Une requête SPARQL (le langage de requête RDF) consistant à rechercher tous les lexèmes descendant de l'étymon FEW *examen* trouvera donc ALW *zamté*. Pour affiner l'analyse, il est possible d'ajouter une deuxième relation indiquant le paragraphe exact du FEW où devrait se trouver ALW *zamté*. Comme nous sommes dans le WEB sémantique, cette description ne consiste pas à associer ALW *zamté* à un numéro de page ou de paragraphe, mais bien à lui associer la catégorie de type « groupement morphologique » qui aura été identifiée comme objet RDF dans la modélisation de l'article EXAMEN du FEW. Cette relation nouvelle entre *zamté* et les critères de regroupement de l'article EXAMEN du FEW permet d'accéder à la mise à jour que fait l'ALW au FEW.

5. Conclusion

En pratique, la formalisation RDF des ressources lexicographiques constitue une couche supplémentaire par rapport aux autres formalisations informatiques apportées à cette ressource. Il est donc tout à fait possible de profiter d'un formalisme XML pour créer de façon automatisée une formalisation RDF, comme cela a été fait pour le DEAF (cf. Tittel / Chiarcos 2018). Il est aussi possible de créer une couche RDF sans passer par une formalisation XML. Les solutions dépendent du degré d'informatisation des ressources que l'on veut intégrer dans le WEB sémantique, ainsi que des objectifs de cette intégration. En fonction de ces objectifs, rappelons que la couche RDF ne doit pas nécessairement représenter l'ensemble des informations présentes dans la structure d'un ouvrage lexicographique. Seules les informations pertinentes pour la mise en réseau des ressources au sein du WEB sémantique sont à prendre en compte.

Dès lors, il est envisageable – et souhaitable – que les ouvrages de référence en linguistique romane fassent l'objet d'une représentation RDF minimale mettant à disposition leur nomenclature et leurs unités lexicales. Le FEW, notamment, pourrait devenir une ontologie précieuse pour toute ressource RDF qui voudrait mettre ses données en relation avec les informations étymologiques contenues dans ce dictionnaire. Il n'est pas nécessaire

pour cela de formaliser l'ensemble des structures complexes de l'ouvrage : la mise à disposition sous un format RDF de sa nomenclature (l'ensemble des étymons vedettes, munis de quelques propriétés dont leur appartenance linguistique et leur 'adresse FEW') rendrait déjà de grands services. La même remarque vaut pour le LEI ou le DÉRom. Cette formalisation RDF minimale permettrait la mise en réseau des lexèmes du domaine roman et, indirectement, l'accès aux nombreuses mises à jour qui sont régulièrement apportées aux ouvrages concernés, tout en conservant l'intégrité des ressources originales.

Université de Lille

Pascale RENDERS

6. Références bibliographiques

- Boutier, Marie-Guy, 2008. « Cinq relations de base pour traiter la matière géolinguistique: Réflexions à partir de l'expérience de l'*Atlas linguistique de la Wallonie* », *Estudis Romànics* 30, 301-310.
- Büchi, Eva, 1996. *Les structures du Französisches Etymologisches Wörterbuch. Recherches métalexigraphiques et métalexicologiques*, Tübingen, Max Niemeyer Verlag.
- Buchi, Éva / Renders, Pascale, 2013. « 46. Gallo-romance I: Historical and etymological lexicography », in : Gouws, Rufus H. *et al.* (éd.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*, Berlin/New York, De Gruyter, Handbooks of Linguistics and Communication Science (HSK) 5/4, 653-662.
- Declerck, Thierry / Wandl-Vogt, Evelyne / Mörrth, Karlheinz, 2015. « Towards a Pan European Lexicography by Means of Linked Open Data », in : Kosem, Iztok / Jakubiček, Miloš / Kallas, Jelen / Krek, Simon (ed.), *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex-2015), electronic lexicography in the 21st century: Linking lexical data in the digital age, August 11-13, Herstmonceux Castle, United Kingdom, Ljubljana/Brighton, Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.*, 342-355.
- DÉRom = Buchi, Éva / Schweickard, Wolfgang (ed.), 2008-. *Dictionnaire Étymologique Roman (DÉRom)*, Nancy, ATILF, <<http://www.atilf.fr/DERom>>.
- Khan, Fahad, 2018. « Towards the representation of Etymological and Diachronical Lexical Data on the Semantic Web », in: McCrae, John P. / Chiarcos, Christian / Declerck, Thierry / Gracia, Jorge / Klimek, Bettina (ed.), *Proceedings of the LREC 2018 Workshop "6th Workshop on Linked Data in Linguistic (LDL-2018)"*, 57-62.
- Khan, Fahad / Boschetti, Federico, 2018. « Towards a Representation of Citations in Linked Data Lexical Resources », in: Krek, Simon / Cibej, Jaka / Gorjanc, Vojko / Kosem, Iztok (ed.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Ljubljana University Press, 137-147.
- Mazziotta, Nicolas, 2011. « L'informatisation du *Französisches Etymologisches Wörterbuch*. Concepts pour une approche modélisée commune à l'*Atlas Linguistique de la Wallonie* », *Zeitschrift für Romanische Philologie* 127, 36-62.

- McCrae, John *et al.*, 2008. *The Lemon Cookbook*, <<https://lemon-model.net/lemon-cookbook.pdf>>.
- Renders, Pascale, 2015. *L'informatisation du Französisches Etymologisches Wörterbuch. Modélisation d'un discours étymologique*, Strasbourg, ÉLiPhi (Travaux de Linguistique Romane - Linguistique de corpus et philologie informatique).
- Renders, Pascale / Baiwir, Esther / Dethier, Gérard, 2015. « Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information », in: Kosem, Iztok / Jakubiček, Miloš / Kallas, Jelen / Krek, Simon (ed.), *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex-2015), electronic lexicography in the 21st century: Linking lexical data in the digital age, August 11-13, Herstmonceux Castle, United Kingdom*, Ljubljana/Brighton, Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 452-460.
- Tittel, Sabine / Chiarcos, Christian, 2018. « Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon », in: Kernerman, Ilan / Krek, Simon (ed.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018), 7-12 May 2018, Miyazaki, Japan*, Paris, European Language Resources Association (ELRA), 58-66.